社会技術研究論文集　Vol.4, 66-74, Dec. 2006

Web マイニングを用いた因果ネットワークの自動構築手法の開発

ASSESSING THE PLAUSIBILITY OF INFERENCE BASED ON AUTOMATED CONSTRUCTION OF CAUSAL NETWORKS USING WEB-MINING

佐藤　岳文[1]・堀田　昌英[2]

[1]修士（工学）ニイウス株式会社　プロジェクト第五担当　(E-mail: U100153@niws.co.jp)
[2]PhD 東京大学大学院助教授　工学系研究科社会基盤学専攻　(E-mail: horita@ken-mgt.t.u-tokyo.ac.jp)

因果ネットワークは，社会的事象の因果関係を系統的かつ視覚的に把握するためのツールとして，さまざまな分野で用いられている．しかし，多くの場合において因果ネットワークの作成は分析者による多大な解釈的作業を必要とし，その構築には多くの時間を要する．本研究では，Web 上にある膨大な文書データを利用して因果ネットワークを自動的に構築するツールを開発し，その活用方法を具体的事例と共に示した．本手法を活用することにより，日常の政策論議の信頼性を，既存の因果的言説を通して検証することが可能になる．

キーワード：因果ネットワーク　推論システム　自然言語処理　政策論議

1. 本研究の背景・目的

社会問題の解決には様々な事象間の因果関係を明らかにすることがしばしば重要となる．しかし多くの事象が複雑に絡み合って起きる社会現象の全体像を把握することはきわめて難しい．各事象間の因果関係を実証することはもちろん，どのような要素的事象が着目している社会問題に影響を及ぼしていそうかさえ容易に判断できないことも多い．

因果ネットワーク(Causal Networks)は，諸事象間の因果関係を有向グラフとして表すことによってその全体像を可視化する手法である．因果ネットワークによって，ある主題に関する様々な因果関係を系統的に把握することができる．また，社会的に受容できない事象を防ぐためにはどのような事象を制御すればよいのかといった，問題解決型のアプローチを採用する際にも因果ネットワークは有用である．

しかし因果ネットワークの作成は多くの場合において分析者の膨大な解釈的作業が必要となる．また，一般にひとつの事象を取り巻く因果関係は非常に複雑であり，因果ネットワークを形成する要素事象は膨大な数に上る．そのため，日々生じる種々の社会問題に関する因果ネットワークを手作業で短期間に構築することは非常に困難である．

本研究では，以上の問題を解決するために web 上にある膨大な文書データを活用して因果ネットワークを自動的に構築するツールを開発した．このツールは，文書の記述から因果関係のデータを自動的に抽出する因果関係自動抽出器と，そのデータを基にネットワークを構築し，それを可視化するとともに，様々な事象の生起を数値的に表す因果ネットワークのツールによって構成される．

2, 3 章でそのアルゴリズムについて述べ，4 章ではこのツールの活用法を実装例と共に示す．最後に 5 章で結論を述べる．

2. 因果関係自動抽出器

因果関係の推論の蓋然性を評価するためには，判断材料となる知識が必要となる．因果ネットワークにおいてその知識は，原因から結果に至る連鎖の間に存在する，中間事象として表現できる．

本研究では，佐藤ら[1)]の手法を基に，web 上にある文書データから，所与の原因と結果の間に位置する中間事象を自動的に抽出するシステムを開発した．

2.1. アルゴリズム

因果関係自動抽出器のアルゴリズムは以下のとおりである．

1．任意の html 文書中の複文・長文を分解し，テキストを単文の集合とする．
2．手がかり標識[2), 3)]によって因果関係が認められた単文のペアについて，それぞれの文章から重要と思われる単語を抽出し，それぞれ原因・結果の事象データとする．
3．因果関係の強さを表す数値と，２で得られた原因・結果二つの事象データを併せたものを，因果関係の１ユニットとして出力する．

(1)　単文への分割

日本語の文章では，因果関係の記述は「～なので～」のように複文の中に現れることが多い．そのため，複文や重文を単文に分割する必要がある．

分割は，あらかじめプログラムに設定する手がかり標識をもとに行われる．手がかり標識は単文の文末に現れる表現であり，これによってプログラムに分割すべき点を知らせる．

まず，文章を形態素解析プログラムMeCab[4)]に入力し，出力された一つ一つの形態素と，それに続くもう一つの形態素を調べる．連続する二つの形態素が本プログラムで規定した条件を満たしたとき，そこは単文の文末と解釈され，分割される．

たとえば，MeCab において，「た　タ　た　助動詞　特殊・タ　基本形」という出力結果は，「書いた」など，動詞，形容詞などの過去形の末尾に現れる形態素を示す．

この次に「ため」という形態素が続いたとき，この文は「～(し)たため，～」という記述だと解釈される．一つの形態素だけでも切れ目を検出することは可能だが，その直前の形態素との関係を調べることで，精度を上げることができる．

手掛かり標識は，「～たら」「～れば」のような順接の因果関係を表すもの，「～だが」「～ものの」などの逆接のもの，「～つつ」「～のち」などの並列のものがある．

この分析過程を適用した例を下記に示す．入力文は，以下のものである．

発車時刻の掲示板を見てから，横須賀線ホームへ行ったが，東海道線上りホームを見ると，多数の人がホームで列車を待っており，何かアナウンスが始まったため，東海道線が先に発車すると思い東海道線上りホームへ移動した[5)]．

この入力文を単文に区切ると，Table 1 のようになる．

Table 1 文章の単文への分割結果

発車時刻の掲示板を見てから,	0
横須賀線ホームへ行ったが,	－1
東海道線上りホームを見ると,	0
多数の人がホームで列車を待っており,	0
何かアナウンスが始まったため,	1
東海道線が先に発車すると思い	0
東海道線上りホームへ移動した.	0

各単文の右に示した数値が，因果関係の有無である．それぞれ 1 を順接，－1 を逆接，0 を並列の因果関係と定義した．

(2)　モダリティ

前節(1)の方法で区切られた一つ一つの単文は各々一つの事象を表していると考えられる．そして，単文の文末の表現によって，因果関係の有無が判別される．

さらに，因果関係の記述から事象間の結びつきの強さを調べることも可能である．例えば，「Aならば必ずB」というように，強調されて記述されている場合，AとBの間の因果関係は強いと解釈できる．

そこで，形態素解析の結果から，モダリティを表しているものを検出する．Table 2 のような設定ファイルを用意しモダリティの手がかり標識を検出する．また，否定の意味を持つ形態素も標識として扱う．これは，助動詞の「特殊・ナイ」と，「特殊・ヌ」，形容詞の「ない」「無い」である．モダリティの強さを表す数値は任意であり，分析者によって自由に設定が可能である．数値設定の結果に対する効果については本研究の対象としない．

Table 2　形態素とそのモダリティ対応表

1.5:	必ず	副詞-助詞類接続
1.2:	きっと	副詞-一般
0.5:	たぶん	副詞-一般
0.5:	だろう	助動詞　特殊・ダ
0.3:	かも	助詞-副助詞
1.5:	とても	副詞-助詞類接続
2:	きわめて	副詞-一般
2:	非常に	副詞-助詞類接続
0.4:	少々	副詞-助詞類接続
0.5:	あまり	副詞-助詞類接続
0.5:	さほど	副詞-一般
2:	決して	副詞-一般

原因－結果の関係にある２つの形態素が検出されたとき，この因果関係の強さを次式 1)で定義する．

（因果関係の強さ）＝（順接（１）or 逆接（－１））
×｛(結果のモダリティ）÷（原因のモダリティ)｝
(1)

因果関係は順接のとき正，逆接のとき負，並列のとき 0 の値を取る．式 1)では，小さい原因で大きな結果が生まれるとき，その因果関係は強いと考える．逆の場合は弱くなる．たとえば，「太郎が叱ったので，次郎はひどく泣き出した」という場合，叱った強さは不明だが，次郎は容易に泣いているので，叱ることと泣くことの間には密接な因果関係がある．逆に，「太郎が強く叱ったので，次郎は泣き出した」という場合，次郎が泣くためには強く叱ることが必要であり，普通の強さで叱っても次郎が泣くかは不明である．そのため，因果関係はやや弱まる．したがって，因果関係は結果が強調されるほど強く，原因が強調されるほど弱くなる．

(3) 各単文の分析

原因，結果それぞれの形態素間の関係を抽出した後，次に各単文を因果ネットワークに組み込む上で適切な形式に変換する．これは同一の事象を言い表す際に用いられる表現のゆらぎや多様性を標準化するためである．

ここでは，それぞれの単文から重要と思われる単語のみを抽出し，単語の集まりとして形式化した．文章の冗長な部分を排し，文を単一の品詞の集合に置き換えることで各単文間の比較が容易になる．

まず，単文の形態素列を構文解析器CaboCha[6]に入力する．CaboChaは，文章を文節に分解するとともに，その係り受け関係を分析するツールである．

CaboCha の出力をもとに，分解された一つ一つの文節について，その性質を分析していく．文節は，主辞と機能語を持つ．主辞はその文節の意味を表す言葉で，主に名詞，動詞，形容詞，形容動詞語幹のいずれかがその役割を果たす．機能語は，文節の文章内における機能を表し，格助詞や係助詞，助動詞などがそれにあたる．

Table 3 はこれを基に「時間の面でもアクセスの面でもＪＲを使うほうが大変便利であったため，」という単文の文節を分析した結果の例である．

たとえば，「使う方が」という文節は，主辞が「使う」であり，機能語が「が」である．主辞は動詞だが，「方」で名詞化しており，文脈の中ではガ格の名詞として機能する．

次に，重要単語の抽出を行う．ここでは，構文解析で得られた係り受け構造から，重要な文節と冗長な文節を判別する．ある文節が重要だと判断されたとき，その文節（被修飾節）の品詞と，それに係っている文節（修飾節）の機能を調べる．

Table 3 文節の分析結果の例

ID	文節	主辞	品詞	機能	係り先
1	時間の面でも	時間	名詞	副詞	2
2	アクセスの面でも	アクセス	名詞	副詞	6
3	JR を	JR	名詞	ヲ格	4
4	使う方が	使う	動詞	ガ格	6
5	大変	大変	名詞	副詞	6
6	便利であったため	便利	形容	－	－

被修飾節の品詞によって，それを修飾する文節としての適切な機能は異なるが，この情報は一般に格フレームと呼ばれる．本システムでは，格フレームを Table 4 のように定めた．

Table 4　本システムにおける格フレーム

被修飾節の品詞	修飾節の機能
名詞	名詞（連体，ト格），動詞（連体形），形容詞
動詞	名詞（ガ格，ヲ格，デ格，ニ格，へ格），副詞
形容（動）詞	名詞（ガ格）

ここでは，まず，単文の最後の文節は述語であると前提し，重要節とする．次に，「重要な文節を，格フレームに則って修飾している文節も重要」という考え方に基づき，係り受け構造をさかのぼって重要節を判別していく[7]．そして，重要節の主辞にあたる形態素を抽出し，その集合を事象データとする．こうして，各単文は標準化されたフォーマットのデータに読み替えられる．

Table 4 で用いた例文の事象データは，「ＪＲ，使う，便利」である．

以上の方法で出力された２つの事象データ及びその間の関係の強さによって一つの因果関係のデータが形成され，それらを統合することによって因果ネットワークを構築することができる．

3.　因果ネットワーク

因果ネットワークのひとつの効用は，ひとたび原因－結果の関係にある事象を重み付き有向グラフとして定式化すれば，ある事象の生起が他の事象にどのような影響を及ぼしうるかを推定できる点にある．本論文では前章で構築された因果ネットワークを用いて，事象の連鎖反応のシミュレーションを行う．またこのシミュレーショ

ン結果と実際の Web 上の記述内容を比較することによって両者の整合度を検証する手法を提案する．

3.1. アルゴリズム

(1) 事象間の因果関係・共起関係

因果ネットワークは，因果関係のデータが与えられたとき，原因の事象を表すノードおよび結果のノードと，それらを結ぶ有向アークを形成する．

一つのノードは一つの事象を表し，前章の因果関係抽出器から得られた単語群をその単位とする．アークは，因果関係の強さを表す値を含んでいる．因果関係の強さは，因果関係抽出器でモダリティの検出によって得られた値である．たとえば，Table.2 より「P が起きると，Q が必ず起きる」という場合は，アークの重みは大きくなり(1.5)，「P が起きると，Q は起きないかもしれない」という場合は，アークの重みは負で，絶対値の小さい値(–0.3)となる．

更に，新たに形成された事象は，既存の事象と比較され，互いに類似したものであれば，それらは双方向アークで結ばれる．これは互いの事象の共起関係を表したものであり，Rが生起することは，Sが生起することと等価であるという意味である．このアークも重みの値を保持しており，それぞれの事象の共起関係の強さを表している．共起関係は，文書に明確に書かれていたものではないので結びつきの強さが比較的弱いものとなる．その計算方法は，両者の事象データである単語群を比較し，単純に一致した割合とした[9),10)]．

Fig. 1 は，事象と，それらの関係を視覚的に示した図である．この例では「『釣り日和である』(P)ならばきっと『水位は上がっていない』(Q_1)だろう」という因果関係の結果事象Q_1が「水位が高い」(Q_2)という事象と 0.50 の共起関係にあるという結果が示されている．

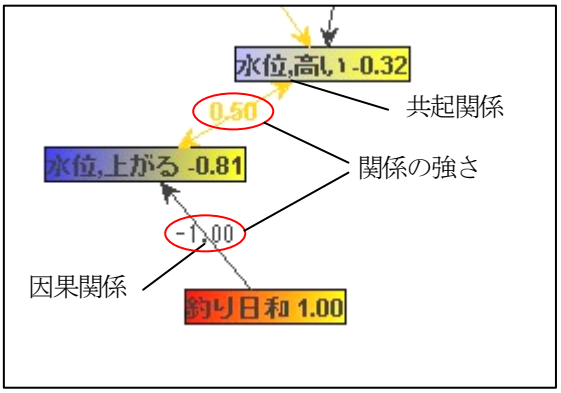

Fig.1 ノードとアークの例

(2) 事象間伝達の減衰関数

因果関係は，ある事象の生起が，他の事象に影響を与えることである．この関係を，本研究では原因の事象が結果の事象に及ぼす影響の度合いとして定式化する．すなわち，原因事象がある定義域内の状態変数 $x \in [-1,1]$ によって記述できるとき，結果事象は x と同じ変域を持つ状態変数 $y=f(x)$の度合いで起きるとする．たとえばxをある年の降雨量の度合い，y をその年の台風被害の頻度などとして解釈できるものとする．

関数fの条件としては，正の因果関係であれば原因が多く生起すればするほど結果も強く生起すること，また，原因が生じても結果は必ずしも起きるとは限らないことを考慮し，減衰関数を用いることが多い[8)]．

本システムでは関数 f にシグモイド(sigmoid)関数を用いた．シグモイド関数は主にニューラルネットワークの伝達関数として用いられており，$f(x)=\dfrac{1}{1+e^{-x}}$ で表され，$x\to\infty$で 1，$x\to-\infty$で 0 に収束する．これを **3.1**(3)で後述する本システムの変域についての要件にあわせ，$x\to-\infty$で－1 に収束するように修正すると，式(2)を得る(Fig.2)．

$$f(x)=\frac{2}{1+e^{-2x}}-1 \qquad (2)$$

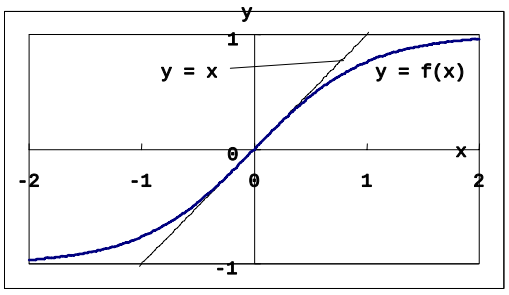

Fig.2　本システムで用いたシグモイド関数

（3）事象生起のシミュレーション

因果関係の知識が蓄積され，ネットワークを形成したのち，それを用いて擬似的に事象を生起させるシミュレーションを行い，各事象間の因果関係の強さを測る．そして，各ページに実際に記述されている因果関係の強さが，シミュレーションの結果とどの程度一致しているかを計算し，各ページの信頼性を導き出す．

このシミュレーションは，最初に一つの事象P_0の強さを表すx_0をランダムに発生させ，連鎖的にその生起を伝達し，他の事象が生起している度合いを測り，P_0と，それぞれの事象間の因果関係の強さを計算するというものである．

その手順は以下の通りである．

まず，P_0に任意の値$x_0 \in [-1,1]$ を入力する．これは，シミュレーションにおいて，一連の連鎖生起の発端が起きたことを意味する．また，P_0と類似したノード，つまりP_0と共起関係にあるノードにも，初期値$x_0 \times w$を入力する．ただしwは**3.1**(1)で定義したP_0との共起関係の強さである．

連鎖生起は，P_0及びその共起事象が起きたとしてその情報を因果関係のある他の事象に伝達していくが，その伝達を経て，その影響がP_0に戻って来ることが考えられる．すると，P_0の状態変数x_0を初期値から更新する必要が生じる．本研究では簡素化のため，各ノードに変化率という係数を導入することによって収束計算を行わないアルゴリズムを用いた．これは，各ノードが他のノードから因果関係の影響を受けたときに，どの程度状態変数の値が変化するかを示す値である．

まず，P_0もしくはP_0の共起事象のいずれでもない事象の変化率を 1 とし，その初期値を 0 とする．次に，P_0は他事象の状態変数の値が伝達されてきてもその影響を受けないこととし，変化率を 0 とする．P_0と双方向アークで結ばれた共起事象については，共起関係のアークをP_0から順次たどっていくことによって，変化率を設定していく．変化率は次式(3)で与えられる．

$$C_R = 1 - \{(1 - C_Q) \times w_{QR}\} \qquad (3)$$

ただし：

C_R　：事象Rの変化率

C_Q　：事象Qの変化率

w_{QR}　：QR間の共起関係の強さ

式(3)より，Qの変化率が低く，Qとの類似度が高い（w_{QR}が大きい）ほど，Rの変化率も低くなる．変化率の範囲は[0,1]である．

P_0およびP_0の共起事象に与えられた初期値をもとに，各事象の生起のシミュレーションを行う．各事象の状態変数の計算方法は次式のとおりである．

$$x_i = \tilde{x}_i + f\left(\sum_k w_{ki} x_k\right) \times c_i \qquad (4)$$

ただし：

x_i　：事象iの状態変数

$\tilde{x}_i$　：x_iの初期値

k　：事象iの原因事象または共起事象

c_i　：事象iの変化率

w_{ki}　：事象kと事象i間のモダリティまたは共起関係の強さ．

この計算をすべての事象に対して行うことで，P_0が生起したときの各事象の状態変数の値を求め，その値から事象間の因果関係の強さを求めることができる．式(2)～(4)より事象ノードの状態変数は全て[-1,1] の範囲の値を取る．P_0の状態変数の初期値がx_0，Qの状態変数の値がx_qだったとき，P_0が原因でQが結果である因果関係の強さは，x_q / x_0となる．

以上の手順を，P_0に指定する事象を変えて行うことで，全ての事象間の因果関係の強さを得ることができる．

3.2. 出力結果

因果ネットワークにおけるシミュレーションの結果，因果関係抽出器で得られたすべての事象の相互の因果関係の強さが得られる．その一例を Fig. 3 に示す．分析結果は Fig. 3 の因果関係マトリクスとして出力される．

	政府,まと	波乱含み	消費税引	参院,控え	慣例
政府,まとめる,歳入一体改革	1.00	−0.76	0.00	0.00	0.00
波乱含み	0.00	1.00	0.00	0.00	0.00
消費税引き上げ法案,提出,	0.00	0.00	1.00	−0.77	0.00
参院,控える	0.00	0.00	0.00	1.00	0.00
慣例	−0.15	0.15	0.00	0.00	1.00
政府,出す,歳出歳入一体改	0.35	−0.34	0.00	0.00	0.00
関税収入,譲渡,止まる	0.00	0.00	0.00	0.00	0.00
自治政府職員,給与財源,不	0.00	0.00	0.00	0.00	0.00
デフレ,つく	0.00	0.00	0.00	0.00	0.00
脱却,展望,至る	0.00	0.00	0.00	0.00	0.00
年度影響試算,示される	0.00	0.00	0.00	0.00	0.00
財政制度等審議,おく	0.00	0.00	0.00	0.00	0.00

結果

原因

Fig. 3　因果関係マトリクス

因果関係マトリクスは，列の事象と，行の事象が対応する位置に，前者を原因とし，後者を結果とする因果関係の強さを記したものである．たとえば，{政府，まとめる，歳入一体改革} と，{波乱含み} との関係は，前者が起きると–0.76 の強さで後者が起き，その逆の因果関係はないということになる．

Fig. 3 を例として出力結果を見ると，事象データとして採用された単語群から文意を十分に把握できるものとそうでないものが混在していることが分かる．本来重要な事象データを精度良く抽出できていれば，例えば {参院，法案提出，控える} という単語群から容易に文を構築できるように，単語群からでも元の事象を一意に定められることが多い．しかしながら現時点における本システムの精度はこの点において十分でない．たとえば {慣例} という単語から元の事象を推定するのは困難であり，さらに他の文章から得られた {慣例} という事象データと同一と見なすことにも問題がある．これは重要語抽出あるいは自動要約にかかわる今後の課題である．

次に，このマトリクスの値と，各ページにおける記述との**相互一致度**をその記述の信頼性にかかわる指標として出力することが可能である．例を Fig. 4 に示す．Fig. 4 は，各 URL で検出された因果関係の記述について，因果関係抽出器で得られた因果関係データ（原因事象，結

URL	原因	結果	関係	相互一致度
www.nikke	ハイテク株も	倉庫株の上	−1	0.10
www.nikke	達する場面も	伸び悩む	−2	0.00
www.nikke	下げる場面も	切り返す	−1	0.01
www.daily.	処理能力が	年内にも引き	−1	0.00
rd.nikkei.c	速報値は4%	上昇基調が	−0.5	0.10
www.asahi	設計や工事	名義貸しをし	1	0.00
www.nikke	約定件数が	システム処理	1	0.00
f1.racing-l	イタリア国内	フィジケラに	−1	0.21
rd.nikkei.c	約定件数が	システム処理	1	0.10
www.373n	緩む	上空に寒気	−1	0.00

Fig. 4　各記述のシミュレーション結果との相互一致度

果事象，および因果関係のモダリティ）と，シミュレーションから得られた因果関係の強さを比較して相互一致度を求めたものである．

相互一致度の定義は次の通りである．あるページに「事象Pが生起すると事象Qはw_{pq}の強さ(モダリティ)で生起する」という推論があったとする．P, Qの状態変数をそれぞれx_p, x_qとおき，因果関係抽出器により求められたP, Qを含む因果関係の集合（アークの集合）をΩとする．ΩはPとQを直接結ぶアークのみからなるΩ_{PQ}とそれ以外のアークの集合Ω^-に分割できる$(\Omega = \Omega_{PQ} \oplus \Omega^-)$．このとき，原因事象$P$の状態変数$x_p$と，因果関係の集合$\Omega$が与えられたときの$x_q$の値を$x_q(\Omega, x_p)$と表すとき，この記述の相互一致度は式(5)で与えられる．

$$D_{PQ} = E\left[\frac{x_q(\Omega^-, x)}{x_q(\Omega_{PQ}, x)}\right] \tag{5}$$

ここで原因事象Pの状態変数$x \in [-1,1]$は確率変数であり，$E[\cdot]$はその期待値である．式(5)は式(4)を用いて下記のように書き換えられる．

$$D_{PQ} = E\left[\frac{x_q(\Omega^-, x)}{f(w_{PQ}x)}\right] \tag{6}$$

D_{PQ}の値はPの状態変数の値をランダムに発生させ，シミュレーションを行うことによって求められる．より簡便化した方法としては，ある記述においてPが完全に生起したと解釈し，$x_p = 1$の値のみを用いて得られるD_{PQ}の値を採用する方法も考えられる．本論文で示されている結果はこの簡便な方法によっている．

本研究では記述の信頼性を，記述された推論が他の推論にどの程度支持されているかという観点から相互一致度という指標を用いて算出したものである．すなわち，ある記述それ自体が推定する結果事象の状態(分母部分)と，当該の記述部分（Ω_{PQ}）を除いた因果ネットワークから推定される結果事象の状態(分子部分)の比を相互一致度と定義している．換言すれば，因果ネットワークのシミュレーションにおいて，ある記述に書かれた通りのことがどれほど頻繁に起きたかを表している．

相互一致度が 1.0 のとき，あるページに記された $P \rightarrow Q$ という因果関係のモダリティは，他のページで検出された「P から Q に至る全ての間接的経路」の合成的モダリティーと一致する．このとき当該の記述は極めて平均的であるということもできる．逆に相互一致度が 1 より大きいとき，この記述は他と比べて因果関係を相対的に弱く表現していることに，0 から 1 の範囲のとき強く表現しているということを意味する．相互一致度が負であることは因果関係の正負が記述とシミュレーション結果で逆であることを示す．

4.　本システムの活用

因果関係抽出器と因果ネットワークからなる本システムは，様々な事象の間の因果関係を把握し，その蓋然性を知る上で有効なツールになる．

4.1.　本システムへのインプット

因果関係に関する推論の評価を行う上で，ユーザーによるシステムへの入力は，以下の３つが考えられる．

1．　キーワード　…　因果関係を把握したいとユーザーが考える主題を表す単語(群)を入力し，その主題に関連する事象全般の因果ネットワークを出力する．
2．　文書データ・ＵＲＬ　…　信頼性を知りたいと思う文書データ（議事録など）を入力し，それに含まれる記述内容の信頼性（相互一致度）を出力する．
3．　文章　…　原因は「風が吹く」，結果は「桶屋が儲かる」というように，それぞれを文章で記述し，信頼性（相互一致度）を出力する．

本システムでは，これら３通りの入力が可能なユーザーインターフェースを構築した．

4.2.　因果関係の視覚化

本システムでは，因果ネットワークをJAVAによるグラフ描画のコンポーネント“JGraph”を用いて可視化の実装を行った[11]．

因果関係抽出器による分析や，ネットワークの構築と平行してグラフィックも描画されるので，Fig. 5 のよう

にネットワークの構築過程を把握することができる.

ユーザーは，因果ネットワークが構築されていく様子を見て，十分な知識が蓄積されたと判断したときに，因果関係抽出器に終了合図を送り，データの収集を止めることができる.

着目すべきノードをクリックすると, Fig. 6 のように，その事象を原因とするシミュレーションの計算結果を表示する．文字の色が反転しているノードが，論の原因となっている事象であり，値は 1.0 となる．その影響を受け, 状態変数が変化した事象は, ノードの色が変化する．より確からしく起きるものは赤くなり，排反事象は青く変化する．また，色の濃さは数値の変化を表している.

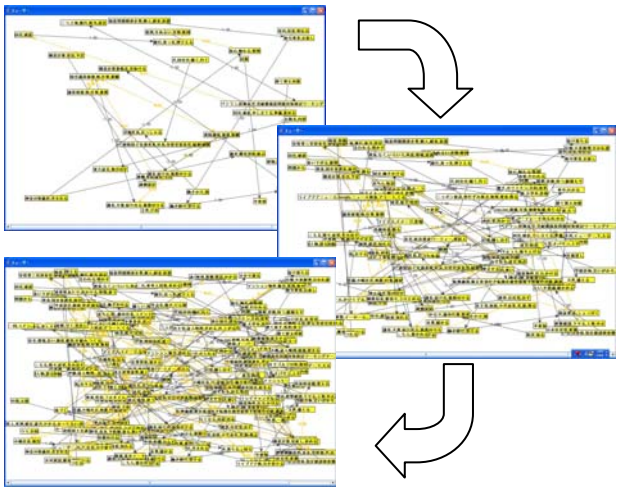

Fig.5　因果ネットワークの構築過程

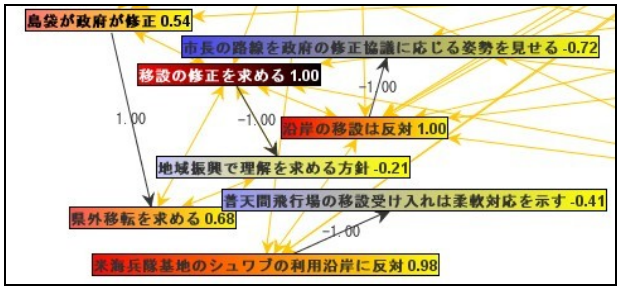

Fig. 6 ノードをクリックした時の出力

ノードを右クリックすると，Fig. 7 のように，因果関係抽出器におけるその事象の分析結果を表示する．その単文における係り受け関係や，各文節の品詞や機能がツリー状に示される.

アークを選択すると，Fig. 8 のようにその因果関係を含む web ページが表示される.
どのようなページからその因果関係を見つけてきたか，実際にどのような文脈の中で因果関係が現れているかを確認することができる.

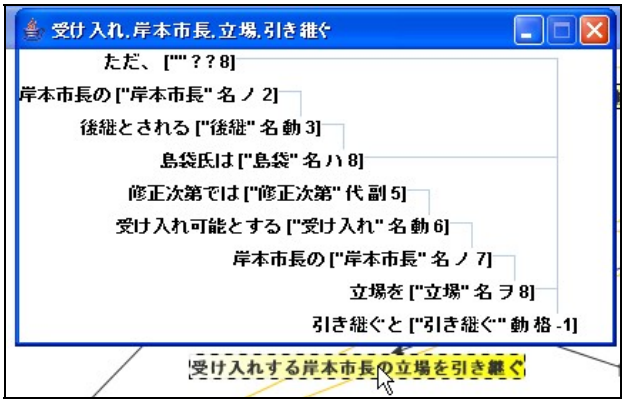

Fig. 7　ノードを右クリックした時の出力

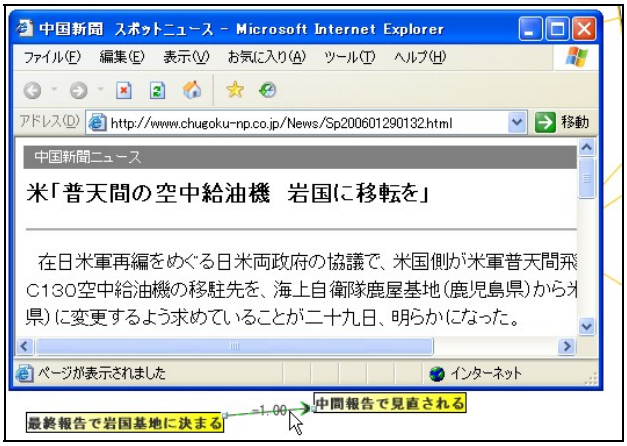

Fig. 8 アークを選択した時の出力

出力結果の妥当性については，現時点で実用化に耐えうる精度を示しているとはいえない. 主な理由としては，①**3.2.**で論じた事象データ抽出の精度が不十分であること，②派生的に主題と無関係な因果的記述をノイズとして取り込んでしまうことの 2 点である．本研究から得られる任意の因果ネットワークのうち，意味判断からして妥当と解釈できる内容の割合は未だ低い．精度向上と妥当性の系統的な検証は今後の研究に譲るが，本研究では相互一致度が比較的高いとされた記述の中から抽出の成功例と考えられるものを下記に例示にするにとどめる.

[入力：キーワード“議員年金”]

- 「首相が通常国会での採決に踏み込む」
  →「郵政法案の二の舞」
- 「政府提出法案の採決は党議拘束をかける」
  →「反対論が根強い法案で党議拘束をかける」
  →「造反議員が出る」

[入力：推論“米国産牛肉の輸入を再開すると狂牛病になる”]

- 「米国産牛肉の輸入を再開する」
  →（途中経路略）→「米国で狂牛病が発生」

≒「狂牛病になる」（相互一致度 0．81）
（ただし「→」は因果関係を，「≒」は共起関係を表す．）

4.3. 情報源の評価

情報が豊富にある昨今において，単体の情報だけでなく情報源自体を評価することの重要性が高まっている．本システムはそのような目的に対しても活用可能である．例えば所与の議題について各報道機関からの記述を集め，各記述の信頼性を前章の定義に従って求めることが可能である．各記述の相互一致度の平均値をその報道機関の代表値とすることで，報道機関の評価をすることができる．

Table 5 は 2006 年耐震強度偽装事件を例にとり，ニュースサイトにおけるこの事件に関する報道の相互一致度を分析した結果である．評価結果自体の妥当性については本研究の対象ではないので報道機関名は匿名で示した．参考のため，各報道機関のホームページ（トップページ）の，GoogleによるPageRank[12]の数値を併せて示した．

Google の PageRank と本システムによる評価との間には以下のような違いがある．前者は，ページに対するユーザーの注目度と密接に関係しており，高くランク付けされたページはさらにランクが上がりやすい再帰的構造を持っている．それに対し後者は，純粋に平均的な内容の記事が高くランク付けされる．したがって本システムは，他との整合性が高い情報もしくは情報源を探索するためのツールと位置づけることもできる．

Table 5　ニュースサイトにおける報道記事の信頼性

メディア名	相互一致度	PageRank
新聞 A	0.96	7 / 10
新聞 B	0.82	6 / 10
新聞 C	0.79	7 / 10
新聞 D	0.77	7 / 10
新聞 E	0.76	6 / 10

5.　まとめ

本研究ではインターネット上の大量の文書から自動的に社会的関心事の因果関係に関する知識を収集する手法を提案・開発した．また実際に解析アルゴリズムとユーザーインターフェースを実装することによって，ユーザーが関心を持った事象の因果関係を対話的に構築することが可能になった．本システムはテキストマイニングの新しい手法として，情報源評価や政策論議の促進等，広範な分野への適用が期待される．

今後の課題としては，インターネット上のノイズ情報の適切な処理方法の構築，既存テキストメディアのインターフェイスとの融合があげられる．

参考文献

1) 佐藤 浩史，笠原 要，松澤 和光：「テキスト上の表層的因果知識の獲得とその応用」信学技報（TL98-23），1999
2) 乾 孝司，乾 健太郎，松本 裕治：「接続助詞「ため」を含む複文から因果関係知識を獲得する」情報処理学会自然言語処理研究会, NL-150-25, 2002
3) 乾 孝司，奥村 学：「文書内に現れる因果関係の出現特性調査」計量国語学, Vol.25, No.3, 2005
4) ＭｅＣａｂホームページ：「MeCab: Yet Another Part-of-Speech and Morphological Analyzer」 http://chasen.org/~taku/software/mecab/
5) http://www.geocities.jp/yuutama_1/830boutou.html
6) ＣａｂｏＣｈａホームページ：「CaboCha: Yet Another Japanese Dependency Structure Analyzer」 http://chasen.org/~taku/software/cabocha/
7) 高橋 哲朗，乾 健太郎，松本 裕治：「テキストの構文的類似度の評価方法について」情報処理学会自然言語処理研究会，NL-150-24，2002
8) 松村 雅明，松本 忠，茂呂 征一郎：「推論システムの構造最適化のファジィペトリネットモデルによる試み」信学技報，CST-2001-41，2002
9) 川本 真司：「嗜好情報獲得のための知識構造化手法」広島市立大学 情報科学部 知能情報システム工学科 修士論文，2001
10) 後藤 将志，大囿 忠親，新谷 虎松：「シソーラスを用いた情報間類似性評価手法について」第 64 回情報処理学会全国大会論文集，pp.73-74，2002
11) JGraph ホームページ：「Java Graph Visualization and Layout」 *http://www.jgraph.com/*
12) 馬場肇：「Google の秘密 - PageRank 徹底解説」 http://www.kusastro.kyoto-u.ac.jp/~baba/wais/pagerank.html

謝辞

本論文の初稿に対して匿名の査読者から多くの有益なご助言を戴いた．ここに記して謝意を表する．

ASSESSING THE PLAUSIBILITY OF INFERENCE BASED ON AUTOMATED CONSTRUCTION OF CAUSAL NETWORKS USING WEB-MINING

Takefumi Sato[1], and Masahide Horita[2]

[1]M.Eng.　Project V Division, Niws Corp., (E-mail: U100153@niws.co.jp)
[2]PhD　Associate Professor, Department of Civil Engineering, University of Tokyo (E-mail: horita@ken-mgt.t.u-tokyo.ac.jp)

Causal networks have been utilized as a tool for structuring complex social phenomena systematically and visually. However, in many cases creating a causal network necessitates a great deal of interpretive works by analyzers. In this study, we have developed a new method for automating the creation of causal networks by utilizing text data on the Web. Examples of its application are illustrated using real cases. It is argued that the proposed method provides ways for verifying the validity of daily life policy discourse through existing causal statements.

Key Words: *causal networks, reasoning systems, natural language processing, policy discourse*